DOCUMENT RESUME

ED 324 016 IR 053 291

AUTHOR Lesk, Michael
TITLE Image Formats for Preservation and Access. A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access.
INSTITUTION Commission on Preservation and Access, Washington, DC.
PUB DATE Jul 90
NOTE 15p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Access to Information; Archives; Comparative Analysis; *Computer Storage Devices; *Information Storage; *Library Materials; Magnetic Tapes; *Microforms; Optical Data Disks; Paper (Material); *Preservation
IDENTIFIERS Deacidification; *Digital Imagery

ABSTRACT

The rapid growth and distribution of scholarly research in the middle and late twentieth century, the limited supply of old books and other paper-based materials, and the deterioration of items printed on acidic paper since the mid-1800s have meant that many libraries lack suitable copies of printed sources that their users would like to read. For some time, libraries have converted books, journals, and newspapers to forms that are more stable, easier and cheaper to copy, and more compact. The most important such form has been microfilm, which is considered a safe, durable, and inexpensive preservation option. Digital imagery is now seen as a viable alternative that offers long-term promise, and is rapidly becoming more accessible to libraries. This report compares digital and microform imagery and emphasizes that making either kind of copy is preferable to leaving acidic paper to decay. Topics addressed include: (1) preservation alternatives, including chemical deacidification, microform, digital imagery, and ASCII (non-image); (2) storage considerations--i.e., magnetic disk, optical WORM (write-once-read-many) disk, digital videotape, digital audiotape, conventional magnetic tape, CD-ROM, magnetic-optical erasable disk, and digital paper; (3) conversion considerations; and (4) transmission considerations. It is concluded that, because microfilm to digital image conversion is going to be relatively straightforward, and the primary cost of either microfilming or digital scanning is in selecting the book, handling it, and turning the pages, librarians should use either method as they can manage, expecting to convert to digital form over the next decade. (SD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED324016

THE COMMISSION ON PRESERVATION AND ACCESS

REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

IMAGE FORMATS FOR PRESERVATION AND ACCESS

A report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access

by

MICHAEL LESK

July 1990

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
Maxine Sitts
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

1785 Massachusetts Avenue, N.W., Suite 313, Washington, D.C. 20036 • (202) 483-7474

The Commission on Preservation and Access was established in 1986 to foster and support collaboration among libraries and allied organizations in order to ensure the preservation of the published and documentary record in all formats and to provide enhanced access to scholarly information.

The Commission is accessible via ALANET (ALA2624) and BITNET (CPA@GWUVM, SITTS@GWUVM), and by FAX (202) 483-6410.

IR053291

Published by
The Commission on Preservation and Access
1785 Massachusetts Avenue, NW, Suite 313
Washington, DC 20036

July 1990

Additional copies available at no charge while supplies last.

This publication has been submitted to
the ERIC Clearinghouse on Information Resources.

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences-Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

© COPYRIGHT 1990 by The Commission on Preservation and Access. Copying without fee is permitted provided that copies are not made or distributed for direct commercial advantage and credit to the source is given. Abstracting with credit is permitted. To copy otherwise, or to republish, requires a fee and specific permission.

Table of Contents

Committee Preface

Introduction 1

Turn the Pages Once 2

Preservation Alternatives 2
- Chemical Deacidification
- Microfilm
- Digital Imagery
- ASCII (non-image)

Storage Considerations 5
1) Magnetic Disk
2) Optical WORM
3) Digital Video Tape
4) Digital Audio Tape
5) Conventional Magnetic Tape
6) CD-ROM
7) Magneto-Optical Erasable Disk
8) Digital Paper

Conversion Considerations 8

Transmission Considerations 9

Conclusions 10

Committee Preface

The Technology Assessment Advisory Committee (TAAC) is a group of seven representatives of industry, publishing, and academia working in the field of digital technology and its applications in scanning, storage, transmission and printing. The group was charged last year with advising the Commission on applications of electronics for the preservation of and access to deteriorating paper-based materials. New technologies with promise for dealing with aging materials include image scanning, compression, and enhancement, as well as networks, optical character recognition, searching algorithms, printers, and user interfaces. This report is one of a series under development by the committee. As such, it is a technologist's summary of how digital technology applies to preservation problems. Although authored principally by Michael Lesk, the report represents the views of the entire committee. It has been issued to stimulate discussion, and not to answer all questions.

—Rowland Brown, Chair
Technology Assessment Advisory Committee

The opinions expressed in this paper are the personal opinions of the authors and are not the corporate policy of their employers. The Committee expresses its thanks to Lee Jones for many helpful suggestions.

Committee members are: (Chair) Rowland C. W. Brown, President, OCLC (retired); Adam Hodgkin, Managing Director, Chenwell Scientific Publishing Limited; Douglas van Houweling, Vice Provost for Information Technologies, University of Michigan; Michael Lesk, Division Manager, Computer Sciences Research, Bellcore; M. Stuart Lynn, Vice President, Information Technologies, Cornell University; Robert Spinrad, Director, Corporate Technology, Xerox Corporation; and Robert L. Street, Vice President for Information Resources, Stanford University.

© 1989 by Sidney Harris — Einstein Simplified, Rutgers University Press

INTRODUCTION

The rapid growth and distribution of scholarly research in the mid and late twentieth century, the limited supply of old books and other paper-based materials, and the deterioration of items printed on acidic paper since the mid 1800s have meant that many libraries lack suitable copies of printed resources their users would like to read. For some time libraries have been converting books, journals and newspapers to forms that are more stable, easier and cheaper to copy, and more compact. The most important such form has been microfilm, which is a safe, durable and inexpensive preservation option. Digital imagery is now an attractive alternative, offering great long-term promise, and is rapidly becoming more accessible to libraries. This paper compares digital and microfilm imagery and emphasizes that making either kind of copy is preferable to leaving acidic paper to decay. The primary expense of salvaging a book is in the selection process and initial handling, while the cost of later conversion from one modern medium to another is comparatively small.

In 1987 the Librarian of Glasgow University complained to me that he had never been sent the first edition of Tristram Shandy (1759-1767) to which the university had been entitled under eighteenth century copyright deposit rules. Since it is a bit late to write to London and berate the Dodsley brothers, what should he do? What should any librarian needing an old book do? Two major problems confront a librarian seeking a pre-1900 book: durability and scarcity. A book printed from the mid-1800s on is probably made of acid paper, bound in a machine-made case, and very fragile. Even earlier books may be in bad shape since the chemical consequences of paper bleaching were not understood when it was first done around 1810, and by 1830 some paper was already deteriorating. Books made in the eighteenth century or before have more durable paper and binding, but the London stationers did not anticipate the number of U. S. libraries that would want copies of these books two hundred years later, and failed to order adequate press runs. Many nineteenth century books, of course, are also in short supply as well as falling apart.

Paper conservation deals only with the physically deteriorating item, not the supply of copies. Today, most bulk deacidification is in experimental or pilot stages, while page-by-page deacidification is expensive. The alternative of publishing facsimile reprints, such as those made by Arno and Scolar Presses, provides both durability and supply, but only the occasional title has an individual demand that will support a new press run. Thus, librarians have favored microfilming as a way of preserving books and other printed items. Microfilming transforms one or more books into a roll of photographic film that is considerably smaller than the original, and that is easy to copy and thus to distribute to other libraries. Microfilm has a very long life, but needs controlled environments. A machine is needed to read it, and many users dislike it.

Digital imagery, where books are scanned into computer storage, is a promising alternative process. Storing page images of books permits rapid transfer of books from library to library (much simpler and faster than copying microfilm). The images can be displayed or printed, much as film images, although with greater cost today. Additionally, digital imagery permits considerable reprocessing: adjustment of contrast, removal of stains, adjustment of image size, and so on. At present the handling of these images still requires special skills and equipment few libraries possess, but there is rapid technological progress in the design of disk drives, displays, and printing devices. Imaging technology will be within the reach of most libraries within a decade.

Digital imagery also may make possible instant reprints, and a new experiment at Cornell University employing very high speed and quality scanning/printing technology will be addressing the feasibility and cost of such an approach. Microfilming deals with preservation, but not with access beyond the library. Digital transmission, combined with workstations in users' offices and nearby printers, offers an opportunity to deliver preserved material in better ways and to more people. Ideally, we might even be able to pay for preservation with revenues derived from improved access.

TURN THE PAGES ONCE

The practical message for the librarian is that the most expensive parts of most preservation activities are (1) selecting the materials to preserve and (2) turning the pages of the selected book for item-by-item chemical treatment, filming, or digitizing. Whether what is done at each page is to spray alkaline buffering solution, make a microfilm image, or digitally scan, the major cost is the time required to gain access to each page. Thus, each book should be handled only once. Chemical paper preservation done sheet by sheet is expensive, must be done on each copy, and does not help alleviate any scarcity of the book. Bulk deacidification, which does not require page-turning, holds out the promise of lower-cost preservation, but also does not increase the number of copies, leaves the original item in its fragile state (except for experimental processes that claim to strengthen the paper), and is not yet at a full production stage. Microfilming and digital imagery, by contrast, make surrogates for the book that are inexpensive to copy. Moreover, conversion between microfilm and digital imagery is much less expensive than conversion to either form from paper.

PRESERVATION ALTERNATIVES

Chemical Deacidification

Bulk deacidification is promised for perhaps $5 to $10 per book. Unfortunately, most mass deacidification processes are currently in either experimental or pilot stages, and some processes involve potentially hazardous chemicals.* (For more information, see "Technical Considerations in Choosing Mass Deacidification Processes," by Peter Sparks, May 1990, published by the Commission on Preservation and Access). With the possible exception of a new British Library experimental process, deacidification merely arrests deterioration for a while; if the book was already fragile, it remains so. From a collaborative perspective, if there are ten copies of an old book scattered around U.S. research libraries, it is likely to be cheaper to film or scan the best available copy once and then reproduce it, than to deacidify all the copies — even in bulk. In addition, microfilming creates a copying master and a bibliographic entry that provide broad access to the information.

Deacidification also can be done on an item-by-item basis at individual libraries. The cost of page-by-page paper treatment, by spraying a chemical fog on the page, is more than the cost of copying, even for one copy. The costs of these more elaborate preservation techniques, which require disassembly and rebinding of each item, are basically prohibitive for books that do not have high value as artifacts. Paper preservation and individual book conservation, however, are the only technologies that preserve the original book itself. For books with particular intrinsic value to scholars (e.g., those whose size or format is significant, or those whose readers are concerned with the manufacture of books, paper, or type), the original copies are important.

* Some libraries further worry that the chemical odor which attaches to deacidified books will be objectionable to their patrons. Good ventilation, unfortunately, is sometimes in conflict with cheap air-conditioning or with fire safety.

(For further discussion of issues related to books as artifacts, see the reports: "On the Preservation of Books and Documents in Original Form" and "Selection for Preservation of Research Library Materials" — both from the Commission on Preservation and Access.)

Microfilm

The process of microfilming a book costs about 10-15 cents per page, not including the cost of choosing the book to microfilm or paying overhead charges to some central organization. Microfilming normally involves producing a roll film master, even if the final version of the book will be on fiche. Microfiche are not considered a preservation format, but can be produced from preservation roll film as an access medium. Microfiche can provide random access to a particular frame faster than roll film, and fiche reading machines are cheaper than microfilm reading machines, which cost several hundred dollars. Fiche are clearly the medium of choice for a microform book catalog, for example. Unfortunately, many readers dislike both film and fiche.

Microfilm, a photographic process, makes a faithful copy of original printed material, including foxing, waterstaining, dark (browning) pages, unsightly borders due to page edges, and faded ink. The use of high contrast film, which is standard, may help with the faded ink at the cost of aggravating discolorations, making it difficult to reproduce continuous-tone images. The photographic materials used for microfilm are very fine-grain and can reproduce the print quality of the original without serious loss (1000 dots per inch). The process of preservation microfilming involves a series of quality control decisions and procedures that are executed throughout filming and developing of the exposed film. Quality monitoring, to determine the success of the quality control procedures, takes place during inspection of the film after it is developed. Both duplication of microfilm and conversion of microfilm to microfiche can be done fully automatically (as can the reprinting from microfilm to paper if desired). Preservation microfilming (or other preservation techniques) must be done more carefully than work intended for only transitory use; thus costs for other kinds of filming or scanning may not be directly comparable.

Roll microfilm comes in a variety of formats. The most common roll film formats are 16mm cartridge and 35mm roll, although preservation microfilming is done primarily in 35mm roll format. Many librarians prefer 35mm film, which provides a larger image readable with less expensive optics, and also offers a better quality source for reprinting. The larger size 35mm film is also more resistant to damage from oxidation, scratching, abrasion, mold, or fungus, since the same amount of damage will obscure a smaller fraction of the page on the larger film. In general, 16mm cartridges can be handled faster automatically and take less space to store, but they also cost more. Progress in photographic technology (such as the development of finer grain films) is improving the images we can make on 16mm film, however.

Although developments are occurring in the use of color microfilm for preservation purposes, nearly all filming or scanning currently is done in high contrast black and white. The practical limits of this large-scale preservation work mean that books with color content, shaded gray scale illustration, or extremely fine printed detail remain, until color filming or better digital technology is available, prime candidates for preservation in their original form.

Digital Imagery

The cost of digitizing a set of images from a book is within a comparable range to microfilming. As in the case of microfilming, the primary cost is again handling. For example, a 30 page/minute 300 dots per inch (dpi) scanner itself costs $13,000; the major cost is obviously not the amortized scanner cost but the cost of the operator. This speed is for sheet-fed operation, with an 80 page stacker, so that attention is required every few minutes. Unfortunately, for old books it is often impossible to process them quickly through a stacker, since the pages are delicate and must be turned carefully. This means substantially higher operator costs on old material or on material that cannot be cut into separate sheets.

The National Library of Medicine has estimated costs based on experiments with a prototype document conversion system developed in-house. This system is designed for bound volumes, fragile paper and face-up capture. The experiments were conducted with a representative sample of the NLM's collection. The system is a distributed, networked, family of AT-based workstations that do document capture, enhancement, compression, quality control (QC) and final storage on WORM digital optical disks. Conversion costs were estimated for a variety of input conditions and in one typical configuration ranged between 13 and 28 cents per page. For details, see: G.R. Thoma, et al., Document Preservation by Electronic Imaging, Volumes I-III, Technical Report of the Lister Hill National Center for Biomedical Communications, NLM, Bethesda, MD., April 1989 — available from NTIS.

Digital scanning can be done at a variety of scan densities. Roughly speaking, 150 dpi is the lowest scanning density that will yield basically acceptable pages for small print. More commonly, scanning is done at 200, 300 or 400 dpi; higher densities are becoming available. Three hundred dpi corresponds to the resolution of most laser printers and is basically able to produce quite acceptable copies, although not quite up to typographic quality (normally considered to start at 1000 dpi). Higher definition is possible but adds considerably to storage cost, for example, doubling the number of dots per inch produces four times as many bits per page.

A 300 dpi 8.5 x 11 inch page is about 1 Mbyte uncompressed, and if filled with dense print as in some journal issues will compress to perhaps 0.2 Mbyte (remember 1 byte contains 8 bits). More normal books (e.g., 5 x 9 inch pages) would be 0.5 Mbyte uncompressed and would compress to under 0.1 Mbyte. Since a typical book is 300 pages long, if uncompressed, six books would fit in a gigabyte (one gigabyte, or Gbyte, is equal to 1,000 Mbytes). If compressed, perhaps 30 books would fit in a gigabyte. If 200 dpi rather than 300 dpi scanning were used, these numbers would become 12 books per gigabyte uncompressed and 45 books per gigabyte compressed (at higher scanning density, data compression is more efficient).

ASCII (non-image)

In contrast to all procedures that preserve the page or the image of the page are techniques for obtaining a computer-readable version of the text. These produce an ASCII file of the characters on the pages. The words are preserved, but not their exact format and appearance. With an ASCII file, it is possible to search for names, specific terms, phrases or, with suitable software, to do various kinds of subject searches. Information can be located much more quickly using computer searches than by flipping through the book, and the thoroughness of a search using a complete text file can be much more complete than conventional indexes. For much of the material considered for preservation, moreover, there is relatively little indexing available; few of our bibliographic secondary services existed in the nineteenth century. ASCII storage is also much more compact; a page of text that will use a few hundred Kbytes in image form will contain only one to two thousand bytes of ASCII, or 1/100th of the space. Other advantages of ASCII storage include the ability to reformat and reprint whole or partial documents easily; the ability to extract quotations or other subsections of the documents and include them in newer papers;* and the ability to mechanically compare texts. Editing texts for later publication also needs ASCII rather than image storage. More ambitious applications such as feeding the texts to speech synthesizers to be read aloud are also possible; perhaps someday we will even be able to do machine translation into other languages.

ASCII text also can be displayed on a wider variety of equipment and on cheaper equipment, than can images (the "glass teletype" 80x24 character screen display costs perhaps $100 while a quality 1000x1000 pixel display is currently over $1000). Even more important is that ASCII displays can be formatted for the particular screen size or program environment preferred by the user; there is less that can be done to rearrange images for display or printing on different devices. The image quality shown does not reflect any fading or discoloration of the original, but merely the quality of the display system. Unfortunately, display systems using ASCII often provide lower quality than that of an image display system because typographic information is

sometimes discarded as the material is converted. Various groups are working on standards for the representation of typographic markup, usually using the SGML format (standard generalized markup language), which will alleviate this problem once in common use. Saving the markup is also important for applications such as reprinting.

Unfortunately, despite many advertisements of OCR (optical character recognition) programs, it is still rather difficult to go from image to character representation. The programs now on the market are adequately fast (10-50 characters per second) for a job that is relatively easy to read (e.g., clear, uniform text), but they are not accurate or versatile enough to handle non-standard type and faded images that are characteristic of old books. Large text conversion projects are still often rekeying, finding this as economical as OCR followed by enough proofreading to maintain accuracy. OCR may well arrive first as a way of doing indexing, where recognizing half the words may well be useful.

STORAGE CONSIDERATIONS

Although digital storage media are being improved, the length of time for safe storage remains well below that for microfilm when stored under appropriate conditions. Ten to 20 years are the figures quoted for most digital optical storage media, with some mention of 100 years. This compares with claims of 500 years of lifetime for microfilm. Even if digital storage media's lifetime is extended, the means of access to the stored information remains the most serious problem. This is because the technology to read the media often becomes obsolete. Who today has a reader for punched cards, 7-track magnetic tape, or 8-inch floppy disks? A librarian who commits to digital storage must expect to have to copy the data regularly ("refresh" the data) until the technology settles down. Fortunately, the cost of doing so is steadily declining.

In addition, digital storage at this time remains relatively expensive. Remember that we are talking about a few dozen books per gigabyte (1,000 Mbytes). The costs of some kinds of digital storage can be reduced by "demounting" — or moving — them to less expensive storage. However, note that this requires an operator step to access the data. Computer media also have several other problems that are serious for librarians. For example, like books, they often require air-conditioned storage. In addition, it is not possible to tell by visual inspection whether computer media have been ruined.

The possibilities for digital storage, as of April 1990, include:

(1) Magnetic disk, usually of the Winchester variety. The current price is roughly $4000 per gigabyte. Access is fast and all material is online. Either software error or hardware error (such as a disk head crash when the reading head touches the disk surface) can destroy the information on a Winchester disk. Thus it is necessary to maintain a copy on some other medium, but the other medium is usually refreshed regularly and does not need to be permanent. The price of magnetic disks has been dropping by almost half each year or so, and the warranty periods doubling. Considerable advances in capacity are still expected: the advent of perpendicular magnetic recording is expected to increase capacity another factor of ten. The equipment is running continuously and some skilled attention is needed.

* Although it may seem that a large nineteenth century library in machine-readable form could raise undergraduate plagiarism to an entirely new level, it would also be easier to check mechanically for such abuses.

(2) Optical WORM (write-once-read-many) disk. A typical drive costs $10,000 to $20,000 and holds two to six gigabytes per removable cartridge. The cartridge is bulky; typically 12-inch diameter platters are used, mounted in housings roughly an inch thick. They can be dismounted, cost about $200, and are reasonably permanent, with 30 to 100 year lifetimes quoted by the manufacturers. Several different manufacturers produce optical WORM drives, and their cartridge formats are not compatible. It is not clear who is going to win in the marketplace; among the vendors are Maxtor, LMSI and Sony. Technological obsolescence of any specific drive is likely to be far more rapid than physical deterioration. There are "jukeboxes" available that can store more than 100 gigabytes, ranging up to more than 300 gigabytes in in one jukebox. The cost of a jukebox starts at $40,000, but larger ones are more likely to be $100,000 or more. These WORM jukeboxes are mechanically very complex devices, and it is not clear whether they will be successful in the long run.

(3) Digital video tape. One vendor, Exabyte, has adapted 8mm videotape into a digital storage medium. The cartridges cost about $6 and store two gigabytes. To access them, of course, the data must be copied back onto a magnetic disk of some sort. There is only one vendor of the systems, it is not clear whether the format will survive, and it is not very durable.* Thus recopying regularly will be necessary. The drive costs about $5,000 (with interfaces, software, etc; if you can do your own mounting and driver coding, the hardware is about $3,000). It takes about two hours to read through a full cartridge.

(4) Digital audio tape (DAT). Several vendors have announced DAT as a computer storage device. The cartridges hold about one gigabyte, are even smaller than the 8mm video cartridges (DAT uses 4mm tape), and the drives cost about $3,600. Again, the format is experimental and it is not clear which vendors' devices will survive. It also is not clear what the lifetime of the cartridges is, but it is unlikely to be permanent and will probably be shorter than 8mm videotape, because the tape is kept under higher tension. Access is faster than on 8mm video cartridge, another consequence of the higher tension of the cartridge. This format is brand new and not yet suitable for use by those who are not interested in testing new devices. Jukeboxes for DAT tape have been announced and are likely to remain in production because of the demand for them in the audio market. At present DAT cartridges cost $20, but this is certain to come down quickly as the format becomes common for consumer audio entertainment.

(5) Conventional 9-track, ½-inch magnetic tape. The physical mechanisms needed to handle such tape are fairly expensive; a sample high performance drive is priced at $16,000. A reel of tape costs $20 and will hold .15 gigabyte, so the cost is about $120 per gigabyte. Tapes must have air-conditioned storage and must be copied every few years, but at least the format is well established and will survive. The durability is better than 8mm video or DAT.

(6) CD-ROM. The CD was designed as a volume production medium but today a single disk can be made for about $1000. It stores a little over 0.5 gigabyte, and there is now agreement on the format of CD-ROM (the so-called "High Sierra" standard). CD-ROM is long-lived, the reader costs about $500, and the format is in fairly wide use for PC data base access. Unfortunately most vendors package specific search software with the data, often with frustrating limitations (designed partly to enforce the copyright law), and it is rare to find the medium used just for storage. Interfaces to large machines and workstations are rare. It is an attractive medium for distribution purposes, however, since the cost of many disks is low (a few dollars per disk). The manufacturing process is not suitable for small scale work, and thus libraries cannot press such disks themselves; the work must be sent out to a company specializing in CD-ROM production. These companies can perform a variety of services, from the relatively simple tasks of mastering and manufacturing a disk, to the more complex work of designing software and retrieval systems for the information provider. Companies include Silver Platter, Meridian Data Systems, Philips-Dupont Optical, and many others.

* The only experiment I know about is one I did myself. Two Exabyte cartridges placed on my car dashboard in June were unreadable in September (New Jersey climate).

(7) Magneto-optical erasable disk. These disks combine magnetic and optical technology to achieve long life, demountable cartridges, and random access. The capacities are now limited to about 0.6 gigabyte per cartridge (using both sides). Drives cost $5000 and the cartridges are $250 each, but likely to become cheaper. Capacities are increasing steadily, and jukeboxes are available. It is not clear which companies or formats will survive.

(8) Imperial Chemical Industries (United Kingdom) has announced "digital paper," a high-density WORM medium using mylar film that can be provided in various shapes and forms. Extremely high density is promised (double that of CD-ROM) but the entire technology is still experimental, more so than any of the alternatives above. No costs are known.

* * *

Here are the cost numbers more directly, with assumptions of: (a) 3-year life (2-year for magneto-optical), based on expected obsolescence of equipment; and (b) $10 charge to recopy, required once per year per reel for the non-durable media. Note that these prices are per gigabyte and should be divided by ten or so to represent the cost per book. I assumed that only ten copies are made of a CD-ROM; this technology is much more appropriate for larger numbers of copies, but it is not realistic to think that there will be much demand for most of these old books.

Medium	Basic Cost/Gbyte ($)	Copying ($)	Total Cost/Gbyte-year ($)
Magnetic disk	4000	0	1300
WORM	75	0	25
Digital video tape	3	5	6
DAT	20	10	17
9-track tape	120	60	180
CD-ROM	2000	0	70
Magneto-optical	400	0	200

Today digital video tape is clearly cheapest if you can deal with the copying requirements; WORM is cheapest if you cannot. Remember that a gigabyte can hold ten books: thus these costs are comparable to the costs of holding a book. The digital video tape and DAT cartridges are substantially smaller than a book, so that they actually represent cheaper storage than on paper. WORM cartridges are fairly bulky and are probably comparable in storage cost to keeping the same material on paper: The cartridge is larger and harder to handle than a book, but it will hold thirty books or so. For all the storage methods above except Winchester disk, the data are assumed to be held "off-line" (meaning that an operator step may be required to mount them for access). Jukeboxes are an alternative to operators. Whether to use on-line storage in a jukebox or off-line storage will depend on the expected use and costs in particular situations.

In summary, it is difficult for a librarian today to install a digital image library. It requires both expertise in computer systems integration and a substantial amount of money — perhaps $100,000 in capital equipment. Remember you need some equipment for people to use any of these media. There are certainly some libraries doing such work (e.g., the National Agricultural Library and the National Library of Medicine) but it is not something to be bought off the shelf or with small resources. But if we assume that the expertise and the capital investment are available, digital image storage is not more expensive than microfilm. Like microfilm, it saves space compared to paper, and digital technology is improving rapidly. Thus digital storage is an appropriate experiment today for the larger libraries, or for groups of libraries.

CONVERSION CONSIDERATIONS

Although the costs of filming and digital scanning (to bitmapped images) are currently within comparable ranges (i.e., filming between 10-15 cents per page; scanning 13-28 cents per page), rekeying the material costs perhaps $1 to $2 or more per page. This is thus an order of magnitude more expensive than any kind of image capture today. On the other hand, rekeying for ASCII access permits rapid search for any particular item within the text. It is valuable to have machine-readable text for old material, but it is not likely to be justifiable for any book for which a new edition is not economically sensible. For any illustrated book, ASCII conversion still leaves behind the question of what to do with the pictorial or graphical material.

Most users of old material will probably be content with the text, but there are some disciplines that need more. As one example, microfilm and digital imagery can cater to people studying aspects of typography, layout, and other aspects of the appearance of old books. Nothing but physical preservation will suffice for those who study papermaking, binding and so on. However, such users are relatively few in number compared with those who want to read the texts. There is a question as to whether even those who wish to read the texts will prefer images of pages to ASCII; more research is needed on this point. In general ASCII storage preserves the words in the text only, not their appearance, and some users express a need for the appearance.

Digital scanning offers flexibility in processing the images: contrast can be adjusted, and image enhancement techniques can be applied either as the image is scanned, or as part of a post-processing phase. Some techniques (e.g., thresholding to adjust for faint printing) need to be performed as part of the archiving process, since they require extra information such as gray level, which may be expensive to store indefinitely; but other techniques can be done later. This is particularly significant, since the most important post-processing technique would be optical character recognition, and it is not yet practical. If OCR technology makes advances, and it becomes possible to process the digital images and convert them to ASCII, then it would be possible to search the content of the books and to reformat or otherwise re-use the material at a much lower cost than rekeying.

Given that digital technology has not yet settled down to the point where libraries can routinely buy document imaging systems off the shelf for prices they can afford, what might a librarian do? (Sticking one's head in the sand is not an acceptable option.) Perhaps most important is to note that once the problem of turning each page is taken care of, the remaining data conversion problems are relatively cheap. To go from microfilm to digital image, in particular, currently can be done at a rate of 2 seconds per image with a Mekel M400 scanner costing $50,000. Operator intervention is needed only every roll or cartridge (that is, perhaps once an hour). This machine is not yet at a state where personnel unskilled in computers can install it, but the operator may be relatively inexperienced. Assuming that we amortized the machine over 5,000 working hours (about 2.5 years of one shift), it would cost perhaps $20 per hour (counting interest, operators, etc.) to run; since in an hour it can do 1,000 to 2,000 frames easily, the cost per frame to convert from microfilm to digital should be perhaps 1 to 2 cents. Compared to the 13-28 cent per page cost of scanning, this means that using microfilm is a reasonable intermediate step to getting digital imagery.

Converting from digital image to microfilm is also possible, although most computer output microfilm recorders are not designed to do graphic images at high speed. Going to paper from both microfilm and digital image is relatively straightforward, and very high speed printers are being developed. It is not clear what the cost will be; the quality will be limited only by the original image, whether scanned or filmed.

The balance between cooperation and individuality must also be struck. Deacidifying a book does not provide more access to that book outside of the library in which the copy is preserved. However, bulk deacidification may force a transition to cooperative work, since the demands and hazards of the bulk chemical processes make them inappropriate for use on a small scale. Microfilming or scanning are likely to be done as part of some group project, since small libraries, in particular, are not likely to have the funds or expertise to provide and use the most advanced equipment.

TRANSMISSION CONSIDERATIONS

If one library has a copy of a book, how can it be sent to another library? Obviously, the physical copy can be loaned, but this deprives the sending library of the book. Microfilm can be duplicated relatively economically (about $10 per reel). It must still, however, be mailed. The combination of duplication and mailing time means that the recipient may wait weeks for a copy. Digital storage has an edge here. In addition to commercial telecommunications networks, such as AT&T's future ISDN service, the US is developing a nationwide digital network running in the megabit* per second range, with experiments in the gigabits* per second range. Today typical transmission speeds are limited by the end equipment to perhaps 100,000 bytes/second. At this rate, it takes about a thousand seconds (i.e., twenty minutes) to send a book anywhere on the net as digital page images. At present connection to the high speed networks (speeds of 1.5 Mbit*) tends to be charged at a flat fee, in the neighborhood of $50,000 to $100,000 per year; at sufficiently high volume the cost of any individual transmission is negligible. The major research universities are already connected at high speeds.

Low-use institutions are more likely candidates for some kind of lower bit* rate, or dial-up or, temporary access. Today this is relatively difficult to arrange at reasonable speed. Service at 9600 baud is quite slow for transmitting whole books as images (it would take a day; my best guess is a cost of $250 or so). If ISDN provides 64 Kbits/sec* service for $10 per hour transmitting 0.1 gigabyte, one compressed book would cost $50 or so to transmit in image format. Of course, many users might want only portions of a book.

Digital transmission around universities is becoming more and more common, and of course computers are now almost ubiquitous and getting more and more powerful, so that with digital storage it will become possible to send copies directly to the offices of many users. Relatively few people, by contrast, have their own microfilm machines. Laser printers capable of printing pages from either image or ASCII storage are also becoming common, offering the possibility of "print on demand" services both centrally, using high speed machines now under development, and remotely, using the user's own equipment. Many office copier machines now being designed, for example, are scanners followed by printers, and could be used for reprinting from digital images. A variety of experiments are being developed to use digital networks to provide current material, and libraries should seek to join with these efforts, using the same networks to provide material that has been preserved.

* I apologize for the conventions by which storage for computer systems is quoted in bytes while communications systems are measured in bits/second. Remember than 8 bits make 1 byte, although the existence of padding in modems means that 10 transmitted bits make one byte at low speeds.

CONCLUSIONS

Some disciplines that rely highly on images and on the book as an artifact in their research will prefer image storage. In the long run, however, scholars are likely to prefer ASCII storage of text for many of their informational needs. ASCII storage permits searching, copying, and duplicating in much more powerful ways than any image storage. Online catalogs, for example, are replacing microfiche catalogs throughout the United Kingdom, and we see no libraries moving towards fiche for catalogs (unless perhaps they are moving from cards). At present, however, it's too expensive to get to full ASCII; and, for most of the relatively rarely used material considered for preservation, it is likely to remain too expensive to use ASCII until optical character recognition becomes feasible.

Digital image storage is practical today, but requires considerable expertise and capital investment on the part of a library trying to do it. However, digital technology is improving very rapidly, much more so than filming. Certainly investment and research should be directed toward digital storage, particularly towards the development of systems that can be used by ordinary libraries. Microfilm is in a similar price range as digital imagery, but is today more accessible to the conventional research library. Because microfilm to digital image conversion is going to be relatively straightforward, and the primary cost of either microfilming or digital scanning is in selecting the book, handling it, and turning the pages, librarians should use either method as they can manage, expecting to convert to digital form over the next decade. Postponing microfilming because digital is coming is only likely to be frustrating and allow further deterioration of important books.